AVANT-PROPOS

Les corpus en questions

François RASTIER
UMR 7114 et ERTIM (Inalco), Paris

Un beau jour du printemps 2004, une journaliste du CNRS vint me trouver et me demanda de lui parler de l'amour au XXI^e^ siècle. Sur ce sujet éminemment consensuel, le *Journal du CNRS* préparait un dossier interdisciplinaire et l'ouvrage que j'avais dirigé quelques années auparavant, *L'analyse des données textuelles — L'exemple des sentiments dans le roman français (1820-1970)* avait semblé me qualifier pour traiter de cette question.
Conscient de mes obligations statutaires, je m'efforçai de répondre, mais je le fis par la question : « Dans quel corpus ? ». Devant le désarroi qui se peignit sur le visage avenant de mon interlocutrice, je me lançai dans des justifications : pour nous, malheureux linguistes, l'amour n'existait que dans les textes et variait avec les discours, les genres et les auteurs. Ainsi n'avait-il rien de commun dans le roman du XIX^e^ siècle, où *amour* trouve pour antonymes *argent* et *mariage*, et dans la poésie de la même époque, où l'argent et le mariage restent évidemment absents. Faute d'avoir eu la présence d'esprit de constituer un corpus sur l'amour en ce siècle naissant, je dus enfin confesser mon incompétence. Tout cela dut paraître bien décevant et il n'en résulta qu'un maigre entrefilet dont je suis confus de n'avoir gardé aucun souvenir.
Il me parut donc nécessaire d'entreprendre une action de communication, non plus à propos de l'amour, sujet apparemment porteur, mais des corpus en lettres et en sciences humaines. Je proposai à Michel Ballabriga et Pierre Marillaud d'organiser à ce propos un colloque, dont voici à grands traits l'argument.

De nombreuses collectivités sont de longue date engagées dans une réflexion sur la numérisation et l'analyse assistée des documents : outre bien entendu les sciences de l'information, il faut mentionner entre autres l'histoire, la sociologie, la linguistique, l'archéologie, les études littéraires.
La constitution et l'analyse de corpus est en passe de modifier les pratiques voire les théories en lettres et sciences sociales. Toutes les disciplines ont maintenant affaire à des documents numériques, et cela engage pour elles un nouveau rapport à l'empirique. En outre, la numérisation des textes scientifiques eux-mêmes permet un retour réflexif sur leur élaboration et leurs parcours d'interprétation. Les nouveaux modes d'accès aux documents engagent-ils de nouvelles formes d'élaboration des connaissances ?
Les nouvelles initiatives prises au plan national et international peuvent devenir l'occasion et donner les moyens d'un projet fédérateur pour les lettres et les sciences sociales.
Aussi ce colloque ouvert entend-il renforcer des liens et favoriser de nouvelles rencontres d'enseignants et de chercheurs de ces disciplines avec ceux des collectivités de la linguistique de corpus et du document numérique. Sans trop d'égard pour l'objectivisme ordinaire, il traite des problèmes philologiques et herméneutiques que pose le travail sur des corpus numériques en fonction des tâches et des disciplines. Il s'attache par exemple à la typologie des genres et discours, à la description de formes et de fonds sémantiques, au repérage de thèmes, à la caractérisation et à l'évolution de concepts, à l'étude des corrélations contenu/expression.
Au plan pratique, il aborde les questions que posent le recueil, l'établissement, le codage, l'étiquetage, le traitement des corpus et leur édition électronique.

On connaît les travers ordinaires des colloques disciplinaires (vedettariat, meurtre du congénère) et des colloques interdisciplinaires (métadiscours grandiloquent) : pluridisciplinaire sans prétendre mettre en scène une interdisciplinarité sans rivages, celui-ci s'est tenu dans une atmosphère sereine de doute enthousiaste, chacun ayant le souci de présenter sa problématique sans en cacher les limites ni négliger les difficultés liées à la constitution des corpus et à l'interprétation des résultats. Des démonstrations de logiciels ont été assurées ainsi que des initiations aux problématiques propres des différentes disciplines concernées.
Le doute positif relève de l'attitude critique nécessaire à toute problématisation scientifique. Il reçoit ici un contenu nouveau, car avec les corpus numériques, les sciences de la culture trouvent de

nouvelles perspectives épistémologiques et méthodologiques, alors qu'elles se trouvent affrontées à des programmes réductionnistes de naturalisation des cultures.
L'objection classique formulée contre leur scientificité tient au caractère non répétable des événements : comme en sociologie, en ethnologie, en psychologie sociale voire en linguistique de l'oral, la présence même de l'enquêteur modifie la situation, on conclut que les sciences de la culture n'auraient donc pas la possibilité d'identifier des causes déterminantes et donc des lois. Or selon le préjugé scientiste qui sous-tend les programmes de naturalisation, la condition nécessaire de la scientificité reste la formulation de lois causales – qu'il faudrait alors chercher dans les substrats physiologiques, neuronaux ou génétiques (cf. Sperber et « l'épidémiologie des représentations » comme explication globale de la culture).
À la classique dualité induction/déduction des disciplines d'observation, le renouvellement méthodologique favorisé par les corpus numériques engage à substituer le cycle suivant : (i) recueil d'information et production des données ; (ii) élaboration de documents scientifiques ; (iii) traitement instrumenté des corpus ; (iv) interprétation des résultats.
La puissance propre de ce dispositif permet de faire émerger de nouveaux observables inaccessibles autrement : par exemple, la phonostylistique, jadis condamnée à l'intuition, se voit à présent pourvue de moyens d'investigation par les statistiques sur corpus phonétisés. En outre, l'utilisation d'une instrumentation scientifique (analyseurs, étiqueteurs, etc.) participe du processus d'objectivation : les objets culturels ont beau dépendre de leur conditions d'élaboration et d'interprétation, les valeurs qu'ils concrétisent peuvent cependant être objectivées comme des faits.
La linguistique de corpus pourvoit ainsi la linguistique d'un domaine où élaborer des instruments et définir une méthode expérimentale propre : elle ouvre aussi des champs d'application nouveaux et engage un nouveau mode d'articulation entre théorie et pratique. D'une part, alors que la linguistique théorique – sans corpus – portait, en extrapolant quelques observations sur des exemples souvent forgés, des jugements universels sur le langage, la linguistique de corpus, sans renoncer à l'élaboration théorique, en limite la portée aux corpus étudiés, et, sans se satisfaire de la seule démarche déductive, procède par essais et erreurs.
En 1999, Chomsky, auteur d'une grammaire universelle, déclarait que la linguistique de corpus n'existait pas, alors même qu'elle était déjà en plein essor : il signalait par ce petit meurtre symbolique qu'elle restait inconcevable pour la linguistique de fauteuil et qu'une rupture épistémologique était en cours. Elle jouit d'une portée générale : en bref, la recherche part d'une diversité constatée, l'unifie dans le point de vue qui préside à la collection du corpus, éprouve son objectivité par l'investigation instrumentée. L'unité, ou du moins la régularité, sera créditée au système, la diversité irréductible au corpus. Ainsi l'opposition entre l'unité substantielle et l'irrégularité accidentelle peut-elle être dépassée dans la description des normes, dont les plus générales, parmi l'ensemble des corpus étudiés, seront considérées comme propres à la langue.

Sans prétendre tirer un bilan prématuré, il semble que cette situation nouvelle impose une reconception de la dualité entre linguistique de la langue et linguistique de la parole, qu'il est de tradition d'opposer, tant chez Bally que chez Benveniste, tant en linguistique de l'énonciation qu'en pragmatique, alors que chez Saussure elles sont parfaitement complémentaires.
On a trop souvent réduit les langues à des dictionnaires et des grammaires, voire à des syntaxes. Il faut cependant tenir compte, outre du *système*, du *corpu*s (corpus de travail et corpus de référence), de l'*archive* (de la langue historique), enfin des *pratiques sociales* où s'effectuent les activités linguistiques. Pour l'essentiel, une langue repose sur la dualité entre un *système* (condition nécessaire mais non suffisante pour produire et interpréter des textes) et un *corpus* de textes écrits ou oraux[1].
Non contradictoire, la dualité dynamique entre corpus et système constitue la langue dans son histoire. Aussi ne saurait-on assimiler la langue historique à la langue fonctionnelle (celle qui fonctionne ici et maintenant) en négligeant que la langue historique détermine la langue

[1] Dans le corpus d'une langue, les *œuvres* tiennent une place particulière parce qu'elles sont valorisées : par exemple l'italien est certes la langue de Dante ; mais son œuvre est le parangon historique qui a présidé à la formation de la langue italienne en tant que langue de culture.
Plus généralement, bien des expressions, dictons et proverbes renvoient aux poètes, législateurs et historiens d'autrefois : ainsi, en chinois, des expressions en quatre caractères qui fourmillent à l'écrit comme à l'oral.

fonctionnelle dans ses structures et ses contenus. Le corpus sert de médiation entre la langue historique et la langue fonctionnelle, et les textes qui n'appartiennent plus qu'à la langue historique entrent dans l'archive. Soit :

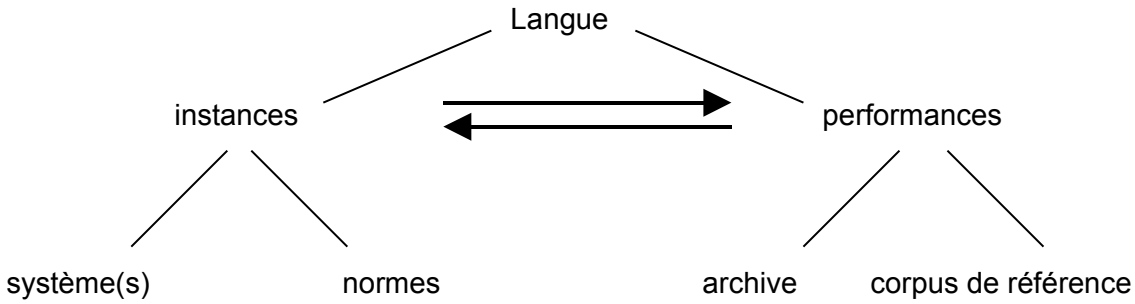
Système(s)	*Corpus*
Langue fonctionnelle	Corpus de référence
Langue historique	Archive

En évoquant le corpus et non les signes, nous soulignons que la langue n'est pas un système de signes – comme le serait un code ; Saussure, à qui l'on prête cette définition, ne l'a jamais formulée. Un signe au demeurant n'a pas de définition intrinsèque : il n'est qu'un *passage*, certes réduit, d'un ou plusieurs textes auxquels il renvoie. Bref, une langue est faite d'un corpus de textes et d'un système. Le système reconstitué par les linguistes est une hypothèse rationnelle formulée à partir des régularités observées dans le corpus.

Entre le corpus et le système, les normes assurent un rôle de médiation : ancrées dans les pratiques sociales, les normes de discours, de genre et de style témoignent de l'incidence des pratiques sociales sur les textes qui en relèvent[1]. Pour éviter la fausse antinomie entre la langue en tant que système de formes et la langue comme produit d'une culture – qui se traduit par la distinction entre cours de grammaire et cours de civilisation – il paraît préférable de considérer que le système comprend des *règles* et des *normes* diversement impératives. Ainsi, les règles de la ballade française font-elles partie du système de la langue française, et elles diffèrent de celles de la ballade anglaise. Entre les règles et les normes, il n'y a sans doute qu'une différence d'évolution diachronique, les règles n'étant que des normes invétérées. En synchronie, toute règle voisine avec des normes qui accompagnent voire conditionnent son application. Ainsi le système d'une langue, à la différence de celui d'un langage formel, est-il en fait pluriel : il se traduit par des régimes structurels différents selon les niveaux et paliers d'analyse. Les domaines d'organisation locaux ou régionaux ne sont pas unifiés dans une hiérarchie attestant l'existence d'un système unique et homogène, comme en témoigne au demeurant l'évolution continue des langues.

Non moins pluriel que le système, le corpus se spécifie *a minima* dans la distinction entre corpus de travail, corpus de référence et archive[2]. À la grande diversité des pratiques sociales correspond celle des corpus produits en leur sein.

Soit, schématiquement :



Enfin, au plan épistémologique, il est vraisemblable que la dualité entre système(s) et corpus traduit une dualité de problématiques, l'une de tradition logique et grammaticale, l'autre de tradition rhétorique et herméneutique[3].

[1] Un texte en effet ne peut pas être produit par un système, comme l'a montré l'échec de la grammaire générative appliquée à des systèmes de génération automatique.

[2] Le *corpus de travail* du linguiste n'est qu'une partie du corpus de référence défini par l'ensemble des textes accessibles dans l'empan spatio-temporel considéré. L'ensemble des performances linguistiques non recueillies sur support constitue le *corpus virtuel* de la langue : il garde une incidence, car toute performance modifie peu ou prou les instances normatives qui lui sont associées (système et/ou normes).

[3] Pour un développement, voir au besoin l'auteur, *Arts et sciences du texte*, Paris, Puf, 2001, introduction.

Problématiques	Logico-grammaticale	Rhétorico-herméneutique
Unités privilégiées	Mot, proposition	Texte
Ordres	Règles	Normes
Sémantique	Signification	Sens
Contextualisation	Minimale	Maximale
Instances	Système(s)	Corpus

La problématique logico-grammaticale privilégie les instances (car elle s'appuie sur une ontologie), alors que la problématique rhétorico-herméneutique privilégie les performances, car elle repose sur une praxéologie.
La sémantique des textes se propose d'articuler les deux problématiques en reconsidérant la première à la lumière de la seconde, car la première peut être obtenue par restriction drastique de la seconde, alors que la seconde ne peut être obtenue par extension de la première. Plutôt donc que de les considérer isolément comme c'est l'usage, il faut tenir compte du fait qu'elles sont modifiées par leur articulation.
Bref, la dualité entre corpus et système(s) n'a rien d'une contradiction : elle est prise dans la dynamique qui constitue la langue dans son histoire et l'institue ainsi en *langue de culture*.
En traitant les corpus, la linguistique renoue nécessairement avec les textes, donc avec la philologie et avec l'herméneutique : la philologie pour les établir et les documenter, l'herméneutique pour les interpréter, y compris dans leur dimension intertextuelle.

L'essor de la linguistique de corpus conduit à préciser le rapport entre textes et documents. Alors que la grammaire travaillait sur l'écrit (son nom même l'indique, littéralement), l'oral est une conquête récente de la linguistique ; encore faut-il qu'il soit fixé sur un support, par enregistrement ou transcription, pour devenir l'objet des débats et conjectures propres à l'investigation scientifique. Textes oraux et écrits trouvent leur première unité dans leur statut de documents.
Plus généralement, les différences entre texte et document, bibliothèque et archive, linguistique de corpus et philologie numérique, sont en train de devenir relatives. Le support numérique ne garantit aucune identité à soi : la restitution de l'inscription est sensible aux formats, aux logiciels de visualisation dont les standards évoluent, si bien que la notion philologique d'herméneutique matérielle doit ici être dépouillée de tout attendu substantiel.
En perdant son unicité, le document numérique se dépouille des qualités du document unique de l'archiviste : authentifiable, doué par sa continuité matérielle d'une intégrité (même quand il est fragmentaire), non reproductible, faisant autorité. L'affichage par pixel détruit toute continuité matérielle qui empêchait les falsifications. Alors qu'une critique initiale suffisait à établir le document, il faut à présent une critique indéfinie pour maintenir une fiabilité. L'établissement des significations doit souvent passer par une succession de versions, dont chacune est le support et le résultat d'une opération de lecture. Changeant de régime, l'objectivation doit être indéfiniment progressive sans pouvoir jamais être considérée comme établie, ce qui engage à rompre avec l'objectivisme pour promouvoir une objectivation critique indéfinie.
Toutefois, ce que le document perd en stabilité, il le gagne en biais d'interrogation. Les logiciels imposent une réflexion théorique sur l'étiquetage, sur les rapports entre méthodes qualitatives et quantitatives : on peut par exemple croiser les résultats de plusieurs méthodes pour faire apparaître de nouveaux observables. C'est autant aux « gens du texte » qu'aux informaticiens de faire des propositions sur ce point : pour aborder ces questions, la voie technologique et la voie épistémologique n'ont rien de contradictoire.
C'est par la méthodologie comparative que l'on va pouvoir exploiter les possibilités techniques actuelles. Pour fonder cette méthode, lui permettre d'évoluer et lui fixer des objectifs de connaissance, il faut aussi que la linguistique assume sa place parmi les sciences de la culture.

La linguistique au demeurant n'a aucune prééminence épistémologique dans la réflexion sur les corpus : l'ensemble des sciences sociales et des disciplines littéraires se doivent d'élaborer à leur propos une réflexion commune en gardant leurs objectifs spécifiques. Elles gagnent à des échanges d'expériences, loin d'une interdisciplinarité fusionnelle d'ailleurs illusoire.
Beaucoup cependant reste à faire pour convaincre de la nécessité de travailler sur corpus. La technicité, le détour instrumental, la notion même de méthode expérimentale, inquiètent certains ; l'attachement à la recherche en fauteuil sans sanctions empiriques, parfois même dans des disciplines littéraires la répugnance à l'égard de toute objectivation censée porter atteinte à la subjectivité souveraine des auteurs et des lecteurs, tout cela conduit certains à considérer l'étude des corpus comme un leurre[1].
Ils formulent une objection récurrente : on ne trouve jamais que ce que l'on cherche. Soit ils regrettent par là que l'on vérifie l'intuition sans songer qu'il est parfois difficile de prouver des évidences, ni que cela fait partie de l'ingrate mission des sciences. Soit ils estiment qu'on trouve toujours quelque chose, c'est faux, car des résultats bruités peuvent inviter au silence. On ne trouve pas toujours ce que l'on cherche, mais parfois autre chose que l'on ne cherchait pas : le corpus, formation heuristique, permet de faire émerger de nouveaux observables.
Certes, on ne trouve trop souvent que ce que l'on sait voir et l'on reste dépendant d'un état de l'art et des problématiques routinières de la « science normale » : il faudra cependant, par une démarche critique, les dépasser ensemble.
Par leurs insuffisances avouées comme par leurs avancées, les études recueillies dans ce livre témoignent d'une recherche ouverte, multidisciplinaire, et se font l'écho des échanges toniques du colloque dont elles gardent mémoire.

Si nous avons beaucoup appris pendant ce colloque, nous le devons aussi au CALS et à ses animateurs : j'ai plaisir au nom de tous les participants à remercier Béatrix et Pierre Marillaud pour leur accueil chaleureux et leur organisation sans faille qui nous ont permis de vivre un moment d'utopie bien présente.

[1] « Dans les programmes nationaux pour les SHS [sciences humaines et sociales], l'accent est souvent mis sur les « terrains », les « corpus » et autres « archives ». Certes, on connaît leur importance pour les SHS, mais on peut douter qu'il s'agisse de priorités scientifiques ou sociétales. On n'imagine pas la chimie des matériaux construire un programme de recherche sur les meilleurs gisements de matières premières ou sur les fournisseurs les plus efficaces en poudres ou autres produits de base ». Fontanille, J. (2006) Supplément d'âme ?, *Vie de la recherche scientifique*, 365, pp. 16-17, ici p. 17. Soit, mais les corpus, nécessairement élaborés par leur constitution même ne sont pas des matières premières. La chimie des matériaux et la besogneuse alchimie des corpus n'ont d'ailleurs pas le même lustre.